University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2020

Optimal Sampling of Parametric Families: Implications for Machine Learning

Huber, Adrian E G ; Anumula, Jithendar ; Liu, Shih-Chii

Abstract: It is well known in machine learning that models trained on a training set generated by a probability distribution function perform far worse on test sets generated by a different probability distribution function. In the limit, it is feasible that a continuum of probability distribution functions might have generated the observed test set data; a desirable property of a learned model in that case is its ability to describe most of the probability distribution functions from the continuum equally well. This requirement naturally leads to sampling methods from the continuum of probability distribution functions that lead to the construction of optimal training sets. We study the sequential prediction of Ornstein-Uhlenbeck processes that form a parametric family. We find empirically that a simple deep network trained on optimally constructed training sets using the methods described in this letter can be robust to changes in the test set distribution.

DOI: https://doi.org/10.1162/neco_a_01251

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-184150
Journal Article
Published Version

Originally published at:
Huber, Adrian E G; Anumula, Jithendar; Liu, Shih-Chii (2020). Optimal Sampling of Parametric Families: Implications for Machine Learning. Neural Computation, 32(1):261-279.
DOI: https://doi.org/10.1162/neco_a_01251

LETTER **Communicated by Meir Feder**

Optimal Sampling of Parametric Families: Implications for Machine Learning

Adrian E. G. Huber
huberad@ethz.ch
Jithendar Anumula
jithendar93@outlook.com
Shih-Chii Liu
shih@ini.ethz.ch
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich 8057, Switzerland

It is well known in machine learning that models trained on a training set generated by a probability distribution function perform far worse on test sets generated by a different probability distribution function. In the limit, it is feasible that a continuum of probability distribution functions might have generated the observed test set data; a desirable property of a learned model in that case is its ability to describe most of the probability distribution functions from the continuum equally well. This requirement naturally leads to sampling methods from the continuum of probability distribution functions that lead to the construction of optimal training sets. We study the sequential prediction of Ornstein-Uhlenbeck processes that form a parametric family. We find empirically that a simple deep network trained on optimally constructed training sets using the methods described in this letter can be robust to changes in the test set distribution.

1 Introduction

The main problems in machine learning are density estimation, regression, and classification based on samples drawn according to an unknown but fixed probability distribution function F. To assess the quality of a machine learner, the notion of generalization was introduced, most prominently in statistical learning theory (Vapnik, 1998, 2013). Statistical learning theory describes conditions on the hypothesis space of the learning algorithm and the number of samples drawn from F such that the empirical risk is close in probability to the expected risk. For generalization to be defined in this framework, it is crucial that the expected risk is calculated with respect to the same probability distribution function that generated the samples used for the evaluation of the empirical risk. A change in the probability distribution function cannot be directly incorporated into statistical learning theory.

Neural Computation **32, 261–279** (2020)
https://doi.org/10.1162/neco_a_01251
© 2019 Massachusetts Institute of Technology

Recent findings have shown, however, that even slight changes in the probability distribution function that generates the data (i.e., different distribution functions for the training or test set) lead to decreases in performance of the learned model (Recht, Roelofs, Schmidt, & Shankar, 2018). This problem can be partially circumvented by including data drawn from different possible probability distribution functions (which are allowed to possess different functional forms) in the training set, effectively demanding that a joint solution is found for all subproblems (Caruana, 1997). In the limit, it is possible that infinitely many probability distribution functions could have generated the data. One possible way of modeling the infinitely many data-generating probability distribution functions is by grouping them into a parametric family.

In this letter, we assume that the data-generating process is itself parametric. Data are then drawn from the whole parametric family: the task that a learning algorithm has to solve is to learn a model for the entire parametric family. Without further prior information on the specific probabilistic structure of the test set, it is a natural requirement to demand that a learned model is equally good for all members of the parametric family. The central question studied in this letter is therefore how training sets containing a finite number of samples can be constructed such that the training set represents the entire parametric family optimally. The tools needed for the analysis carried out in this letter mostly stem from information theory, specifically universal coding theory, and not from machine learning (Rissanen, 2007; Cover & Thomas, 2012).

For the sake of clarity and in order to derive quantitative statements, we focus on a specific stochastic process, the Ornstein-Uhlenbeck process. Being both a gaussian and Markovian process, this stochastic process is rich in structure while still being analytically tractable. Most of the results we present, however, apply to more general problem classes.

The problem of how to optimally sample from a parametric family is tightly connected to universal coding theory. Some universal coding inequalities described in section 2 directly correspond to the problem of sequential prediction in the case of an Ornstein-Uhlenbeck process as shown in section 3. The specific stochastic process chosen therefore yields a task (sequential prediction—having observed a time series up to sample n, sample $n+1$ is predicted) that directly corresponds to questions of how to sample a parametric family optimally in the sense of universal coding theory. The letter concludes by empirically studying the generalization behavior shown by deep networks trained on the Ornstein-Uhlenbeck parametric family in an autoregressive manner. We empirically find that a simple model trained on optimally constructed training sets generalizes better to changes in the test set distribution than if the model is trained on suboptimally generated training sets.

We use the following notation. Let $x^n = (x_1, x_2, \ldots, x_n)$ be a sequence of real-valued elements and $X^n = (X_1, X_2, \ldots, X_n)$ a sequence of random

variables on $\mathbb{R}^n$. In this work, X^n will denote strictly stationary stochastic processes. Define a set of probability density functions (PDF) $\{P_\lambda, \lambda \in \Omega\}$ on $\mathbb{R}^n$ with Ω a compact subset of $\mathbb{R}^m$, assuming there are m free parameters. $|\cdot|$ denotes the operation of taking the determinant of a square matrix. $\log(\cdot)$ is the natural logarithm.

2 Review on Universal Coding

We give a brief description of ideas from the universal coding literature that are crucial for this work. Assume that a family of PDFs $\{P_\lambda, \lambda \in \Omega\}$ on $\mathbb{R}^n$ and an observed sequence $x^n = (x_1, x_2, \ldots, x_n)$ (which is generated by one of the densities in the family) is given. If the specific PDF P_λ generating x^n is known, then the entropy rate $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_\lambda\left[-\log\left(P_\lambda\left(X^n\right)\right)\right] = H(\lambda)$, with $\mathbb{E}_\lambda[\cdot]$ the expectation with respect to P_λ, corresponds to the best compression of the source. Such a compression statement follows from the asymptotic equipartition property (AEP; Cover & Thomas, 2012). For the sampled strictly stationary Ornstein-Uhlenbeck process, which is discussed in section 3 in more detail, the AEP holds (Barron, 1985). If P_λ is not known, however, the question arises of whether it is still possible (asymptotically in n) to reach the entropy rate of the stochastic process, provided that the parametric family $\{P_\lambda, \lambda \in \Omega\}$ is known. Universal coding theory answers this question in the affirmative for a wide class of parametric families (Merhav & Feder, 1998). To show this, a mixture source $P(x^n) = \int_\Omega w(\lambda) \cdot P_\lambda(x^n)\, d\lambda$ is introduced, with w a PDF (we do not consider cases in which w might be discrete) on Ω. This mixture source can then be used as a replacement for the unknown P_λ. A natural question associated with such a mixture source is how w should be chosen. It is intuitively clear that mixture sources $P(x^n)$ set up by different w will behave differently. It turns out that a particular choice of w carries with it a notion of channel capacity. Let Λ denote a random variable with PDF w on Ω. The parameters λ indexing P_λ are realizations of Λ. The prior w^*, which reaches channel capacity $C_n = \sup_w I_w(\Lambda; X^n)$ with channel input Λ and channel output X^n, where $I_w(\Lambda; X^n)$ denotes mutual information induced by $w(\lambda) P_\lambda(x^n)$, maximizes the mutual information between Λ and X^n. If Λ is distributed as w^*, then observations x^n generated by P_λ contain most information about the m parameters in Ω. Additionally, w^* has the further property of being the prior that induces maximin redundancy (Merhav & Feder, 1998). The channel capacity C_n is furthermore a lower bound on the Kullback-Leibler divergence between the true data-generating distribution P_λ and any other PDF $Q(x^n)$ (Merhav & Feder, 1995):

$$D(P_\lambda || Q) > (1 - \epsilon) C_n. \tag{2.1}$$

Inequality 2.1 holds for all $\epsilon > 0$ and for all $\lambda \in \Omega$ except for some λ in a subset $B \subset \Omega$ whose size under w^* vanishes at an exponential rate with

C_n. For $w = w^*$, $D(P_\lambda || P^*) = C_n$, with P^* the mixture source with capacity-achieving prior w^*. Hence, for w^*, nearly all sources P_λ lie on or close to a hypersphere centered at P^* with Kullback-Leibler divergence equal to C_n, as can be inferred from the previous discussion and inequality 2.1. It is crucial to emphasize that this statement holds only for the capacity-achieving prior w^*. Other mixture sources based on different priors w will in general be closer to some subset of sources in the parametric family $\{P_\lambda, \lambda \in \Omega\}$ and have larger Kullback-Leibler divergence than C_n to other sources in the parametric family.

It is interesting to note that for the parametric family introduced in section 3 (sampled strictly stationary Ornstein-Uhlenbeck processes), an asymptotically accurate form of the channel capacity can be deduced (Rissanen, 1996),

$$C_n = \frac{m}{2}\log\left(\frac{n}{2\pi}\right) + \log\int_\Omega \sqrt{|I(\lambda)|}d\lambda + o(1), \tag{2.2}$$

with $o(1)$ tending to zero for $n \to \infty$ and $I(\lambda)$ the Fisher information matrix of the stochastic process,

$$I_{ij}(\lambda_*) = \lim_{n\to\infty}\frac{1}{n}\left\{\frac{\partial^2}{\partial\lambda_i\partial\lambda_j}\mathbb{E}_{\lambda_*}\left[-\log\tilde{P}_\lambda(X^n)\right]\right\}_{\lambda_*}, \tag{2.3}$$

with i and j ranging from 1 to m and λ_* in Ω.

An additional interpretation of C_n can be given in terms of the number of distributions in $\{P_\lambda, \lambda \in \Omega\}$ that are distinguishable based on the observation of a sequence of length n (Balasubramanian, 1996; Rissanen, 2007). It is intuitively clear that different sources in the parametric family $\{P_\lambda, \lambda \in \Omega\}$ are not necessarily distinguishable after observing n samples. This notion can be made more precise by using the language of hypothesis testing. For the parametric family discussed in this letter, this analysis is described in section 3. Note that equation 2.2 is a consequence of choosing Jeffreys' prior in the mixture source $P(x^n)$, which is given by the following expression (Jeffreys, 1998),

$$w_{\text{Jeffreys}}(\lambda) = \frac{\sqrt{|I(\lambda)|}}{\int_\Omega \sqrt{|I(\lambda')|}d\lambda'}, \tag{2.4}$$

which is asymptotically equal to the capacity-achieving prior w^* for the parametric family considered in this letter. The number of distinguishable distributions after observing a sequence of length n is roughly equal to e^{C_n}. Since Jeffreys' prior, equation 2.4, is asymptotically capacity inducing, the maximal number of distinguishable distributions is reached for Jeffreys'

prior. More precisely, if Λ is distributed according to w_{Jeffreys}, then the sampled stochastic processes P_λ are maximally distinguishable on average. Any other prior w would (at least asymptotically) lead to a smaller number of distinguishable distributions. This argument can be strengthened by appealing to the analog of equation 2.1 for arbitrary priors (Merhav & Feder, 1995). It can be shown that $D(P_\lambda||Q)$ is larger than $(1-\epsilon)C_R$, with $\epsilon > 0$ and C_R equal to the logarithm of the maximal number of random sources chosen under the prior w that can be distinguished in the sense of having a bounded error probability (Merhav & Feder, 1995). Q is an arbitrary distribution on x^n as in equation 2.1. The inequality holds again for all parameters λ except in a set $B' \subset \Omega$ whose size measured by w tends to zero for $n \to \infty$ under certain conditions.

The previous ideas, although formulated in terms of probabilities (equivalently, in terms of log-loss) can be directly applied to the case of sequential prediction under the mean squared error (MSE) loss, at least for the Gauss-Markov processes used in this letter. This idea is described in section 3.

3 Lower Bounds on the Sequential Prediction Error

In this section, we first introduce the parametric family studied in this letter. Thereafter, we derive lower bounds on the sequential prediction error under the MSE loss for different priors w from which the strictly stationary sampled Ornstein-Uhlenbeck processes are drawn.

3.1 Some Results on the Ornstein-Uhlenbeck Process. The Ornstein-Uhlenbeck process is defined as

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t, \tag{3.1}$$

with $\theta > 0$, $\mu \in \mathbb{R}$, $t \geq 0$, $\sigma > 0$ and W_t the standard Wiener process. For the process to be strictly stationary, the first value x_0 at time $t = 0$ is drawn from a gaussian distribution with mean μ and variance $\frac{\sigma^2}{2\theta}$. In the strictly stationary case, the Ornstein-Uhlenbeck process can be alternatively written as

$$X_t = \mu + \frac{\sigma}{\sqrt{2\theta}} e^{-\theta t} W_{e^{2\theta t}}, \tag{3.2}$$

with $\{W_{e^{2\theta t}}\}$ a time-scaled Wiener process. We next derive some bounds on the growth of strictly stationary Ornstein-Uhlenbeck processes. These bounds are needed in the explicit construction of the recurrent neural network (RNN) that implements the asymptotically optimal solution of the sequential prediction problem described in section 4.1. To understand the growth behavior of the strictly stationary Ornstein-Uhlenbeck process, the law of the iterated logarithm is invoked:

$$\limsup_{t\to\infty} \frac{|W_t|}{\sqrt{2t\log\left(\log\left(t\right)\right)}} = 1 \quad \text{a.s.} \tag{3.3}$$

By applying the law of the iterated logarithm to the time-scaled Wiener process, the denominator of equation 3.3 is changed to $\sqrt{2e^{2\theta t}\log\left(\log\left(e^{2\theta t}\right)\right)}$, while the numerator is replaced by $|W_{e^{2\theta t}}|$. Multiplying the denominator by $\frac{\sigma}{\sqrt{2\theta}}e^{-\theta t}$, one obtains $\frac{\sigma}{\sqrt{\theta}}\sqrt{\log\left(2\theta t\right)}$. Hence one can conclude the following about the second term of equation 3.2:

$$\limsup_{t\to\infty} \frac{\sigma}{\sqrt{2\theta}} e^{-\theta t}\, |W_{e^{2\theta t}}| = \frac{\sigma}{\sqrt{\theta}}\sqrt{\log\left(2\theta t\right)}. \tag{3.4}$$

For a finite $t > 0$, there will in general exist a constant $C > 0$ such that the strictly stationary Ornstein-Uhlenbeck process in $[0, t]$ will be almost surely contained within the interval

$$\left[\mu - C\frac{\sigma}{\sqrt{\theta}}\sqrt{\log\left(2\theta t\right)}, \mu + C\frac{\sigma}{\sqrt{\theta}}\sqrt{\log\left(2\theta t\right)}\right]. \tag{3.5}$$

3.2 Sampling the Ornstein-Uhlenbeck Process. We consider Ornstein-Uhlenbeck processes drawn from a parametric family. The two free parameters are $\mu \in (c, d)$ with $c, d \in \mathbb{R}$, $d > c$ and $\theta \in (a, b)$ with $a, b \in \mathbb{R}^+$, $b > a$. $\sigma \in \mathbb{R}^+$ is arbitrary but fixed. The uniformly sampled Ornstein-Uhlenbeck process amounts to an autoregressive AR(1)-process,

$$X_{n\delta} = e^{-\theta\delta} X_{(n-1)\delta} + \mu\left(1 - e^{-\theta\delta}\right) + \epsilon_n, \tag{3.6}$$

with $\epsilon_n \sim N\left(0, \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta\delta}\right)\right)$ independent over time, $\delta > 0$ the distance between consecutive samples and $X_{n\delta}$ the nth sample. $(X_\delta, X_{2\delta}, \ldots, X_{n\delta})^\top$ is distributed according to a multivariate normal distribution with mean vector $(\mu, \mu, \ldots, \mu)^\top$ and covariance matrix:

$$\Sigma = \frac{\sigma^2}{2\theta}\begin{pmatrix} 1 & e^{-\theta\delta} & \ldots & e^{-\theta(n-1)\delta} \\ e^{-\theta\delta} & 1 & \ldots & e^{-\theta(n-2)\delta} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\theta(n-1)\delta} & e^{-\theta(n-2)\delta} & \ldots & 1 \end{pmatrix}. \tag{3.7}$$

We next derive the asymptotic Kullback-Leibler divergence between two strictly stationary Ornstein-Uhlenbeck processes as well as the Fisher information matrix of this stochastic process. Both are needed for the subsequent discussion of distinguishability, as well as for the explicit construction of

Jeffreys' prior. The inverse of covariance matrix 3.7 is given by

$$\Sigma^{-1} = \frac{2\theta}{\sigma^2\left(1-e^{-2\theta\delta}\right)} \cdot \begin{pmatrix} 1 & -e^{-\theta\delta} & 0 & \cdots & 0 \\ -e^{-\theta\delta} & 1+e^{-2\theta\delta} & -e^{-\theta\delta} & \cdots & 0 \\ 0 & -e^{-\theta\delta} & 1+e^{-2\theta\delta} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -e^{-\theta\delta} \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \tag{3.8}$$

which is a symmetric tridiagonal matrix. From equation 3.8, we obtain the determinant of Σ,

$$|\Sigma| = \frac{1}{|\Sigma^{-1}|} = \frac{\sigma^{2n}\left(1-e^{-2\theta\delta}\right)^{n-1}}{(2\theta)^n}. \tag{3.9}$$

The asymptotic Kullback-Leibler divergence is then equal to

$$\begin{aligned} D\left(\mu_1,\theta_1||\mu_0,\theta_0\right) &= \lim_{n\to\infty}\frac{1}{n}D\left(P_{(\mu_1,\theta_1)}||P_{(\mu_0,\theta_0)}\right) \\ &= \frac{1}{2}\frac{\theta_0}{\theta_1}\frac{1}{1-e^{-2\theta_0\delta}}\left(1-2e^{-(\theta_0+\theta_1)\delta}+e^{-2\theta_0\delta}\right) \\ &\quad + (\mu_1-\mu_0)^2\frac{\theta_0}{\sigma^2\left(1-e^{-2\theta_0\delta}\right)}\left(1-e^{-\theta_0\delta}\right)^2 \\ &\quad -\frac{1}{2}+\frac{1}{2}\cdot\log\left(\frac{1-e^{-2\theta_0\delta}}{1-e^{-2\theta_1\delta}}\right)+\frac{1}{2}\cdot\log\left(\frac{\theta_1}{\theta_0}\right). \end{aligned} \tag{3.10}$$

Evaluating the Fisher information matrix, equation 2.3, for the strictly stationary sampled Ornstein-Uhlenbeck process, we find

$$I\left(\mu_*,\theta_*\right) = \begin{pmatrix} \frac{\left(\left(e^{2\theta_*\delta}-1\right)-2\theta_*\delta\right)^2}{2\theta_*^2\left(e^{2\theta_*\delta}-1\right)^2}+\delta^2\frac{1}{e^{2\theta_*\delta}-1} & 0 \\ 0 & \frac{2\theta_*}{\sigma^2}\frac{e^{\theta_*\delta}-1}{e^{\theta_*\delta}+1} \end{pmatrix}, \tag{3.11}$$

where we first differentiate with respect to θ and then with respect to μ. Equation 3.10 can be locally approximated as follows:

$$\begin{aligned} &D\left(\mu_1,\theta_1||\mu_0,\theta_0\right) \\ &\quad\approx \frac{1}{2}\begin{pmatrix}\theta_1-\theta_0 & \mu_1-\mu_0\end{pmatrix} I\left(\theta_0,\mu_0\right)\begin{pmatrix}\theta_1-\theta_0 \\ \mu_1-\mu_0\end{pmatrix}. \end{aligned} \tag{3.12}$$

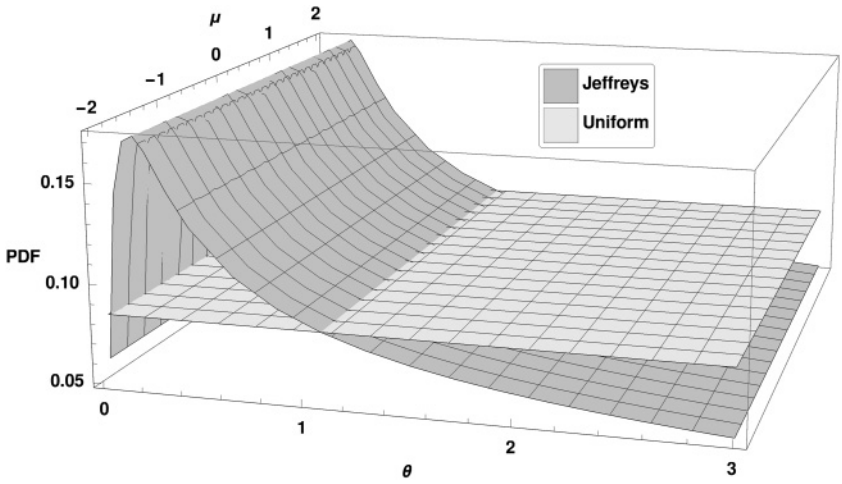

Figure 1: Jeffreys' and uniform prior for the Ornstein-Uhlenbeck process.

Equation 3.12 is a quadratic approximation to equation 3.10; it corresponds to a Taylor expansion truncated after the second expansion coefficient. Equation. 3.11, plugged into equation 2.4, yields Jeffreys' prior for the parametric family composed of sampled Ornstein-Uhlenbeck processes. Jeffreys' prior is shown in Figure 1 for $\delta = 10$.

3.3 Lower Bounds. In section 2, various lower bounds under log loss were discussed that pertain to representing a parametric family by some mixture source. Here we discuss lower bounds under MSE loss for the task of sequential prediction tailored to the sampled Ornstein-Uhlenbeck process.

Theorem 1. *Consider gaussian ARMA processes with compact parameter space* $\Omega \subset \mathbb{R}^m$, $m > 0$, *and* p *autoregressive terms,* $p < m$. *Given any prior* w *on* Ω *with corresponding random coding capacity* C_R *and any* $\epsilon > 0$, *the following lower bound is valid for all parameters* λ *except in a set* $B' \subset \Omega$ *whose size measured by* w *tends to zero for* $n \to \infty$*:*

$$\frac{1}{n-p}\mathbb{E}_\lambda\left[\sum_{i=p+1}^{n}\left(X_i - h_i\left(X^{i-1}\right)\right)^2\right] \geq \sigma^2(\lambda)\left[1 + (1-\epsilon)\frac{2C_R}{n-p}\right], \tag{3.13}$$

with $\sigma^2(\lambda)$ *the variance of the stationary Wold decomposition of the stochastic process and* $\hat{x}_i = h_i\left(x^{i-1}\right)$ *any measurable prediction function.*

Proof. The random coding theorem (Merhav & Feder, 1995) holds for gaussian ARMA processes. In case P_λ and Q from equation 2.1, as well as its extension to the random coding case, are both gaussian distributions, the random coding theorem leads directly to a lower bound on the MSE loss. $P_\lambda(x^n)$ is the probability of data sequence x^n induced by the gaussian ARMA model, while $Q(x^n)$ is obtained by converting the arbitrary prediction function $\hat{x}_i = h_i\left(x^{i-1}\right)$ into a PDF:

$$Q\left(x_i|x^{i-1}\right) = \sqrt{\frac{1}{2\pi\sigma^2(\lambda)}} e^{-\frac{\left(x_i - h_i\left(x^{i-1}\right)\right)^2}{2\sigma^2(\lambda)}}. \tag{3.14}$$

The prediction begins after observing p initial values. We then find that

$$\mathbb{E}_\lambda\left[\log\frac{P_\lambda(X^n)}{Q(X^n)}\right] = -\frac{1}{2}(n-p) + \frac{1}{2\sigma^2(\lambda)}\mathbb{E}_\lambda\left[\sum_{i=p+1}^{n}\left(X_i - h_i\left(X^{i-1}\right)\right)^2\right], \tag{3.15}$$

which, upon rearranging and insertion into the random coding theorem and division by $n-p$, yields equation 3.13. □

Corollary 1. *For a strictly stationary sampled Ornstein-Uhlenbeck process with sampling interval $\delta > 0$, the following lower bound is obtained:*

$$\frac{1}{n-1}\mathbb{E}_{(\mu,\theta)}\left[\sum_{i=2}^{n}\left(X_{i\delta} - h_i\left(X^{(i-1)\delta}\right)\right)^2\right] \geq \frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}\left[1 + (1-\epsilon)\frac{2C_R}{n-1}\right]. \tag{3.16}$$

Proof. By choosing $\sigma^2(\lambda) = \frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}$ and $p = 1$ according to the Ornstein-Uhlenbeck process specifications, equation 3.6, the desired result is obtained. □

Remark 1. If the prior w is chosen as Jeffreys' prior, then the random coding capacity C_R can be replaced by C_n from equation 2.2 in the case of gaussian ARMA processes.

Theorem 1 is a generalization of a well-known lower bound obtained for a uniform prior w (Rissanen, 1984). The greatest lower bound results from choosing Jeffreys' prior. In the case of a uniform prior w, the number of distinguishable distributions is proportional to $n^{\frac{m}{2}}$, provided that some parameter estimators exist that converge sufficiently fast (cf. Merhav & Feder, 1995). The conditions hold for the strictly stationary sampled Ornstein-Uhlenbeck process. In that case, C_R in inequality 3.13, has to be

replaced by $\frac{m}{2}\log(n)$ with $m = 2$ in our case on account of the number of free parameters in the Ornstein-Uhlenbeck parametric family. Note that if w was chosen such that only one distribution could be effectively distinguished, the lower bound would be equal to $\frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}$. The same lower bound would be reached if the two free parameters θ and μ were known and would not have to be estimated first. The second part of equation 3.13 $(1-\epsilon)\frac{2C_R}{n-p}$, hence measures the additional complexity of having unknown free parameters.

The lower bound in equation 3.13 for Jeffreys' prior and the lower bound for the uniform prior can be reached asymptotically. By estimating the AR coefficient $\psi_1 = e^{-\theta\delta}$ and $\psi_2 = \mu\left(1-e^{-\theta\delta}\right)$ with ordinary least squares (OLS), which for the Ornstein-Uhlenbeck process coincides with a maximum likelihood (ML) estimation of the two parameters conditioned on the first observation, and using these estimates to predict the next sample $\hat{x}_{i\delta} = \hat{\psi}_1 x_{(i-1)\delta} + \hat{\psi}_2$, the following error is obtained: $\mathbb{E}_{(\mu,\theta)}\left[\left(X_{i\delta} - \hat{X}_{i\delta}\right)^2\right] = \frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}\left(1+\frac{2}{i}\right) + O(i^{-\frac{3}{2}})$ (Fuller & Hasza, 1981, 1980). Summing the previous expression from $i = 2$ to n and dividing by $n-1$, one obtains

$$\begin{aligned}
&\frac{1}{n-1}\sum_{i=2}^{n}\mathbb{E}_{(\mu,\theta)}\left[\left(X_{i\delta}-\hat{X}_{i\delta}\right)^2\right] \\
&\quad= \frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}\left(1+\frac{2\left(H_n-1\right)}{n-1}\right) \\
&\qquad + O\left(\frac{H_n^{\left(\frac{3}{2}\right)}-1}{n-1}\right),
\end{aligned} \tag{3.17}$$

with H_i being the ith harmonic number and $H_i^{(m)}$ the ith generalized harmonic number. For $n \to \infty$, H_n can be approximated by $\log(n)$, while the second term tends to zero. Hence, the lower bound in equation 3.13 can be reached asymptotically in the Ornstein-Uhlenbeck case, as can be seen by inspecting the asymptotic behavior of the term $\frac{C_n}{n-1}$ with C_n given by equation 2.2.

3.4 Distinguishability of Processes from the Ornstein-Uhlenbeck Parametric Family. We construct explicit regions of indistinguishability for the Ornstein-Uhlenbeck parametric family. If only a finite number of samples are given, then distinct strictly stationary Ornstein-Uhlenbeck processes will not be distinguishable if their parameters (θ_0, μ_0) and (θ_1, μ_1) are too close to one another in a suitable sense. To make this notion more precise, we construct regions of indistinguishability around (θ_0, μ_0) such that, given n samples, the process corresponding to parameters (θ_0, μ_0) and

a process corresponding to parameters drawn from the region of indistinguishability around (θ_0, μ_0) will not be effectively distinguishable. The analysis is based on a related investigation of distinguishability for independent and identically distributed (i.i.d.) stochastic processes (Balasubramanian, 1996). Let us therefore assume that a realization of the random vector $(X_\delta, \ldots, X_{n\delta})^\top$ has been observed. $\mathrm{P}_{(\theta_0,\mu_0)}$ corresponds to the null hypothesis, while $\mathrm{P}_{(\theta_1,\mu_1)}$ is the alternative hypothesis. The observed random vector is drawn from either $\mathrm{P}_{(\theta_0,\mu_0)}$ or $\mathrm{P}_{(\theta_1,\mu_1)}$. Assuming that the type 1 error probability α_n is bounded from above by a constant $\epsilon \in (0, 1)$, $\alpha_n \leq \epsilon$, the minimum type 2 error probability,

$$\beta_n^\epsilon = \inf_{\substack{\mathrm{A}_n \subseteq \mathbb{R}^n \\ \alpha_n \leq \epsilon}} \beta_n, \tag{3.18}$$

with A_n an acceptance region for the null hypothesis, is given asymptotically (via a generalized Stein's lemma; Vajda, 1989) as

$$\lim_{n\to\infty} -\frac{1}{n}\log\left(\beta_n^\epsilon\right) = D\left(\mu_1, \theta_1 || \mu_0, \theta_0\right). \tag{3.19}$$

For a fixed number of samples n, we then find the following region of indistinguishability around (θ_0, μ_0),

$$\begin{aligned} \frac{\kappa}{n} &\geq D\left(\mu_1, \theta_1 || \mu_0, \theta_0\right) \\ &\approx \frac{1}{2}\begin{pmatrix} \theta_1 - \theta_0 & \mu_1 - \mu_0 \end{pmatrix} I\left(\theta_0, \mu_0\right) \begin{pmatrix} \theta_1 - \theta_0 \\ \mu_1 - \mu_0 \end{pmatrix}, \end{aligned} \tag{3.20}$$

with $\kappa = -\log\left(\beta^*\right) + \log\left(1 - \epsilon\right)$ and β^* a constant between 0 and 1. For sufficiently large n, β^* will be smaller than β_n^ϵ, showing that the type 2 error will be greater than a certain constant. Equation 3.20 shows that the regions of indistinguishability around (θ_0, μ_0) are given by ellipses whose major axes depend on the local value of the Fisher information matrix. Starting with such regions of indistinguishability, a covering of parameter space can be carried out. An illustration of such a procedure is given in Figure 2 with parameters $\beta^* = 0.95, \epsilon = 0.01$, and $\delta = 0.1$ for two different sequence lengths, $n = 50$ and $n = 100$.

4 Empirical Results with Deep Networks

The results described in sections 2 and 3 are intrinsic properties of parametric families. We first recapitulated general results of universal coding theory and derived specific results for the Ornstein-Uhlenbeck parametric family thereafter. By an empirical analysis, we show in this section that the

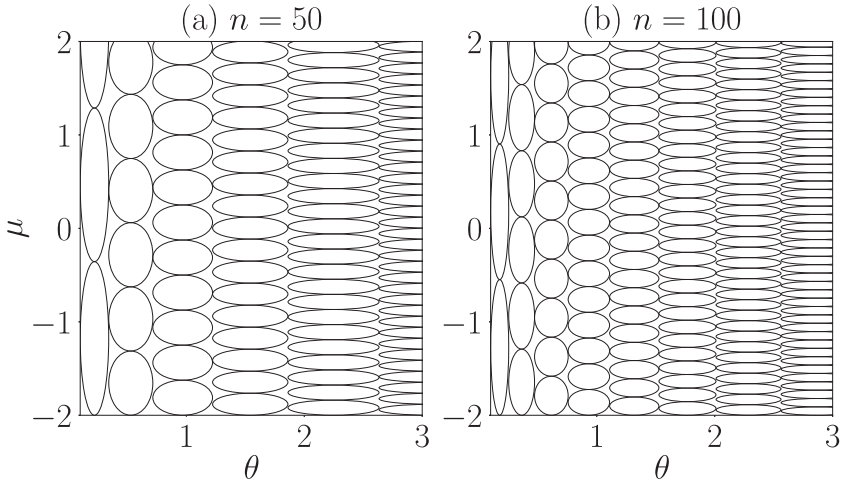

Figure 2: Coarse illustrative partition of parameter space by regions of indistinguishability.

previous statements have repercussions for machine learning as well. The choice of the specific learning algorithm is to some extent arbitrary for this task. We have hence chosen standard RNN architectures with long short-term memory (LSTM units; Hochreiter & Schmidhuber, 1997), as these are state of the art for time series prediction.

We first describe a constructive scheme to approximate the optimal solution from section 3.3 within the hypothesis space of an RNN. The approximation methods described in section 4.1 are used to verify that the chosen RNN architecture described in section 4.2 can in principle approximate closely the optimal solution. To carry out the approximations, the results from equation 3.5 and the appendix are required as the domain of the input to the RNN needs to be known.

4.1 Approximating the Optimal Solution through Explicit Construction. An RNN with a single hidden layer with LSTM units is used for the sequential prediction task. In order to approximate the solution based on the OLS equations discussed in section 3.3 (cf. Fuller & Hasza, 1980, for the OLS equations), each subexpression in the OLS equations is approximated through one of the units in the recurrent layer. In order to approximate the expression $x^2 + y$, for example, we first approximate x and y through two of the recurrent units, x^2 with another unit, and finally $x^2 + y$ with a fourth unit. The OLS equations contain both polynomial terms of second order as well as reciprocal terms.

Three main ideas are used for the approximation of the equations with the LSTM layer. The first idea is to rescale the input to the approximately linear region of the corresponding tanh/sigmoid nonlinearity. This step requires a careful analysis of the growth behavior of the individual terms in the OLS equations. Equation 3.5 provides an upper and lower bound within finite time intervals for the strictly stationary Ornstein-Uhlenbeck process, with $C \approx 1$ from numerical simulations. From this, as well as a more thorough analysis of the growth behavior of terms in the OLS equations detailed in the appendix, it is possible to obtain scaling factors that ensure that the rescaled input is within the linear region for some finite time horizon. The second idea is to approximate the multiplication operation required in the OLS equations by the use of Hadamard multiplication in the LSTM update equation for the cell state. The last idea is to approximate the division operation by first approximating the inverse of the divisor and then using the multiplication approximation to multiply the dividend and the inverse of the divisor. For the approximation of the inverse, we can either train a subnetwork to approximate the operation within our range of interest or we can use a constructive approximation scheme closely based on previous work (Jones, 1990).

4.2 Training on Jeffreys' Prior and Uniform Prior. To elucidate the importance of sampling of the parameter space on the performance of the RNN, we train two networks with the same configuration and training conditions: one where the process parameters are sampled according to Jeffreys' prior and the other where the sampling is carried out according to a uniform prior. We choose a network with a single layer of 100 units, followed by a linear transformation to a single dimension for the prediction. This network can approximate the optimal solution closely. The network is trained with stochastic gradient descent with a learning rate of 0.001 with early stopping. The range of the parameter μ for the process is $(-2, 2)$, while the range for the parameter θ is $(0.01, 3)$. The sampling interval δ is set to 10, while n is arbitrarily set to 500.

Both of the trained models are tested on sequences drawn from the two priors: Jeffreys' and uniform. The results for the case of 50 parameters sampled during training are shown in Table 1. The results are averaged over 5 draws of parameter sampling and 10 random initializations of the network for each draw.

It is observed that with an increasing number of parameter samples drawn from the parameter space, the difference in the performance of the models trained on the two priors gets smaller. This can be seen in Figure 3, in which the performance of the models trained on stochastic process realizations drawn from the two priors (Jeffreys' and uniform) and tested on Jeffreys' prior is plotted against the number of stochastic process realizations drawn.

Table 1: Comparing the Performance (MSE) of Models Trained on the Two Priors and Tested on the Two Priors.

		Test Prior	
		Uniform	Jeffreys'
Train Prior	Uniform	2.91 ± 0.4	3.83 ± 0.25
	Jeffreys'	2.94 ± 0.32	3.4 ± 0.2
	Optimal	0.79	1.11

Note: "Optimal" is related to the lower bounds from section 3.3.

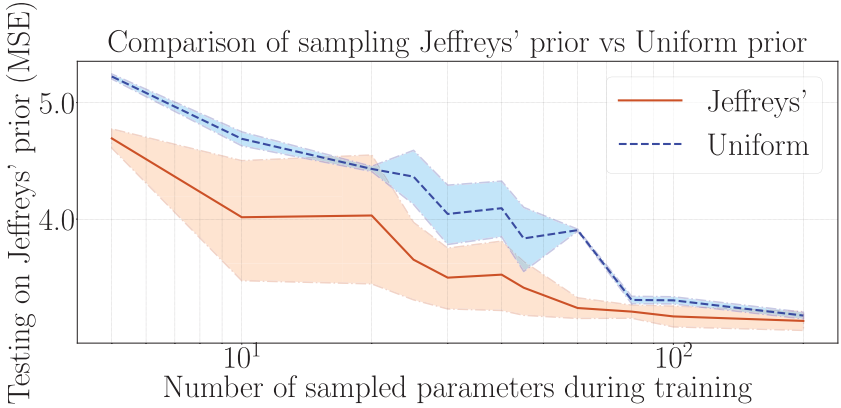

Figure 3: Comparing the performance (MSE) of the models trained on two priors, tested on Jeffreys' prior, with an increasing number of sampled parameters during training.

5 Discussion

Classical machine learning theory investigates the learnability of relationships from i.i.d. samples drawn from a fixed but unknown probability distribution, as alluded to in section 1. For the non-i.i.d. case, extensions of statistical learning theory, type guarantees have been developed (cf. Kuznetsov & Mohri, 2015; McDonald, Shalizi, & Schervish, 2017, as well as references therein). Generalization is always understood to refer to the same distribution generating the training and test set.

If multiple distributions are to be learned, it is natural to require the model to do equally well on all of them. This requirement can be directly translated into the language of universal coding theory. The number of independent realizations of stochastic processes p drawn independently according to some prior w on the compact parameter space, as well as the

length n of each stochastic process realization, are, as is intuitively clear, crucial for any required theory of generalization in the parametric family context. In classical statistical learning theory, n, as well as the complexity of the hypothesis space, is the main focus of investigation. For finite n, only finitely many stochastic processes are distinguishable. Asymptotically in n, for the stochastic processes considered in this letter, the capacity-inducing prior will be given by Jeffreys' prior. Since the maximum number of distinguishable models is close to e^{C_n}, p will have to be at least equal to e^{C_n}. In fact, since C_n is in general growing with increasing n, the minimum number of required stochastic process realizations p will depend on n. The dependence of p on n therefore implicitly reflects the fact that the number of distinguishable distributions in a parametric family grows with increasing n. Since the capacity-inducing prior w^* is the prior under which the maximum number of distributions in the parametric family are distinguishable, it follows that p adapted to this prior is sufficient for any other prior. Finding a p adapted to w^* is therefore a necessary requirement if one attempts to learn the entire parametric family. The empirical counterpart of this statement for the case of MSE loss is found in Figure 3 as well as Table 1. Training on stochastic process realizations drawn from Jeffreys' prior ensures that testing on a different prior (here the uniform prior was chosen) does not lead to an increased MSE loss. Training on the uniform prior and testing on Jeffreys' prior, however, leads to a marked increase in MSE loss.

The capacity used in the lower bound equation 2.1, as well as in the lower bound equation 3.13, is the capacity of the parametric family, not the capacity of the hypothesis space of the machine learner. Notions of capacity for the machine learner reflect the richness of the class of functions that such a learner can approximate. The capacity C_n measures the richness of the parametric family.

Assume that it was known only that a set of observations could be modeled by a parametric family with m free parameters, while the specific form of the parametric family was not known. In such a case, it would not be possible to obtain p such that, uniformly for all possible parametric families with m free parameters, p would be sufficient to guarantee that any parametric family could be fully learned (in the sense that the solution found should be close to a mixture source induced by the capacity-achieving prior). If the form of the parametric family was not known, it seems reasonable to use stochastic process realizations drawn uniformly from the space of parameters. If the capacity-inducing prior, however, was very different from the uniform prior, then most of the obtained realizations from the uniform prior would not facilitate learning the parametric family fully. The ill-adapted sampling mechanism would prohibit an optimal learning of the parametric family. The testing error in Figure 3, with testing performed by drawing stochastic process realizations from Jeffreys' prior and training carried out by using either Jeffreys' or the Uniform prior, converges to the same error for increasing p. This behavior is expected in view of the fact

that the two priors are positive everywhere within the parameter space, as can be seen in Figure 1. A more subtle analysis of this fact can be carried out by noting that the number of distinguishable distributions under both priors is not too different from one another as discussed in section 3.3 for the parametric family considered in this letter.

Equation 3.13 provides a lower bound on the sequential prediction error for the MSE loss, assuming that the form of the parametric family was known. The empirical results obtained in section 4, do not require knowledge of the specific form. By the explicit construction detailed in section 4.1, we show that a solution close to an optimal solution lies in the hypothesis space of the chosen network architecture. It is hence guaranteed that the chosen deep network is in principle well specified. The results shown in Table 1 indicate that the empirical solution found by the network does not reach the lower bounds, here denoted by "Optimal", implying that an inefficiency exists in the optimization procedure. A thorough analysis is outside the scope of this letter, however, as it would require an investigation of the loss landscape of the chosen deep network with stochastic process realizations drawn according to some prior w as input, as well as of the optimization algorithm used.

Empirically, it was observed in the experiments that if one first trains the deep network with observations drawn from some prior w_1 until convergence and thereafter changes the prior to some w_2 and continues training, the previously found solution changes. This behavior is expected in view of the previous discussion, as a changed prior induces a different optimal solution. It follows that there is a close link between optimal solutions and the sampling of parameter space.

Most of the previous statements hold for more general families of distributions and not only for parametric families. Equation 2.1, as well as the statements on the capacity-achieving prior, hold in particular in more general contexts (Merhav & Feder, 1995). The simple form of the capacity, equation 2.2, as well as the fact that Jeffreys' prior is asymptotically capacity inducing are, however, not correct in a more general context. To achieve optimality, however, the sampling mechanism should still be matched to w^*.

Appendix

We derive some results needed for the explicit construction of the RNN used to implement the aymptotically optimal solution for the sequential prediction of the sampled strictly stationary Ornstein-Uhlenbeck process. Let us study the time integral of the strictly stationary Ornstein-Uhlenbeck process:

$$Y_t = \int_0^t X_s ds. \tag{A.1}$$

$\{Y_t\}$ is a gaussian process, implying that it is fully characterized by its mean and covariance function. For the mean as a function of t, one obtains

$$\begin{aligned}\mathbb{E}[Y_t] &= \mathbb{E}\left[\int_0^t \mu + \frac{\sigma}{\sqrt{2\theta}} e^{-\theta s} W_{e^{2\theta s}} ds\right] \\ &= \int_0^t \mathbb{E}\left[\mu + \frac{\sigma}{\sqrt{2\theta}} e^{-\theta s} W_{e^{2\theta s}}\right] ds = \mu t, \end{aligned} \tag{A.2}$$

with the exchange of integration and expectation order justified by Fubini's theorem, while the covariance function is given by

$$\begin{aligned}\mathrm{Cov}(Y_t, Y_s) &= \mathbb{E}[Y_t Y_s] - \mu^2 ts = \mathbb{E}\left[\int_0^s \int_0^t X_a X_b da db\right] - \mu^2 ts \\ &= \frac{\sigma^2}{2\theta^3}\left(e^{-\theta s} + e^{-\theta t} - e^{-\theta|t-s|} + 2\theta \min(s,t) - 1\right). \end{aligned} \tag{A.3}$$

We next analyze the time integral of the squared strictly stationary Ornstein-Uhlenbeck process:

$$Z_t = \int_0^t X_s^2 ds. \tag{A.4}$$

The expectation of $\{Z_t\}$ is given by

$$\begin{aligned}\mathbb{E}[Z_t] &= \mathbb{E}\left[\int_0^t \mu^2 + \sqrt{2}\mu \frac{\sigma}{\sqrt{\theta}} e^{-\theta s} W_{e^{2\theta s}} + \frac{\sigma^2}{2\theta} e^{-2\theta s} W_{e^{2\theta s}}^2 ds\right] \\ &= \left(\mu^2 + \frac{\sigma^2}{2\theta}\right) t, \end{aligned} \tag{A.5}$$

while the covariance function is

$$\begin{aligned}\mathrm{Cov}(Z_t, Z_s) &= \mathbb{E}\left[\int_0^s \int_0^t X_a^2 X_b^2 da db\right] - \left(\mu^2 + \frac{\sigma^2}{2\theta}\right)^2 ts \\ &= \frac{\sigma^4}{8\theta^4}\left(e^{-2\theta s} + e^{-2\theta t} - e^{-2\theta|t-s|} + 4\theta \min(t,s) - 1\right) \\ &\quad + \frac{2\mu^2\sigma^2}{\theta^3}\left(e^{-\theta s} + e^{-\theta t} - e^{-\theta|t-s|} + 2\theta \min(t,s) - 1\right). \end{aligned} \tag{A.6}$$

$\{Z_t\}$ is not a gaussian process. We study sums of the form $\sum_{i=1}^n X_{(i-1)\delta}$ with a sampling interval δ and $\{X_t\}$ a strictly stationary Ornstein-Uhlenbeck process. $\left(X_0, X_\delta, \ldots, X_{(n-1)\delta}\right)^\top$ is distributed according to a multivariate normal distribution with mean vector $(\mu, \mu, \ldots, \mu)^\top$ and covariance matrix:

$$\Sigma = \frac{\sigma^2}{2\theta} \begin{pmatrix} 1 & e^{-\theta\delta} & \dots & e^{-\theta(n-1)\delta} \\ e^{-\theta\delta} & 1 & \dots & e^{-\theta(n-2)\delta} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\theta(n-1)\delta} & e^{-\theta(n-2)\delta} & \dots & 1 \end{pmatrix}. \tag{A.7}$$

Hence it follows that the sum $\sum_{i=1}^{n} X_{(i-1)\delta}$ is distributed according to a gaussian distribution with mean $n\mu$ and variance:

$$\operatorname{var}\left(\sum_{i=1}^{n} X_{(i-1)\delta}\right) = \frac{\sigma^2}{2\theta} \frac{2e^{-\theta(n-1)\delta} - 2e^{\theta\delta} + n\left(e^{2\theta\delta} - 1\right)}{\left(e^{\theta\delta} - 1\right)^2}. \tag{A.8}$$

Next, sums of the form $\sum_{i=1}^{n} X_{(i-1)\delta}^2$ are studied. We find $\mathbb{E}\left[\sum_{i=1}^{n} X_{(i-1)\delta}^2\right] = \left(\mu^2 + \frac{\sigma^2}{2\theta}\right)$ and

$$\begin{aligned} \operatorname{var}\left(\sum_{i=1}^{n} X_{(i-1)\delta}^2\right) = \frac{\sigma^2}{2\theta} \Bigg(& \frac{8\theta\mu^2\left(e^{-\theta(n-1)\delta} - e^{\delta\theta} + ne^{\delta\theta} - n\right)}{\left(e^{2\delta\theta} - 1\right)^2} n\sigma^2 \\ & + \frac{2\sigma^2\left(e^{-2(n-1)\delta\theta} - e^{2\delta\theta} + ne^{2\delta\theta} - n\right)}{\left(e^{2\delta\theta} - 1\right)^2} \\ & + \frac{8\theta\mu^2\left(e^{-\theta(n-1)\delta} - e^{\delta\theta} + ne^{\delta\theta} - n\right)}{\left(e^{2\delta\theta} - 1\right)^2} \Bigg). \end{aligned} \tag{A.9}$$

Given that $\sum_{i=1}^{n} X_{(i-1)\delta}$ is a gaussian random variable, $\left(\sum_{i=1}^{n} X_{(i-1)\delta}\right)^2$ will be a noncentral χ^2 distribution. $\frac{\left(\sum_{i=1}^{n} X_{(i-1)\delta}\right)^2}{\frac{\sigma^2}{2\theta}\frac{2e^{-\theta(n-1)\delta} - 2e^{\theta\delta} + n\left(e^{2\theta\delta} - 1\right)}{\left(e^{\theta\delta} - 1\right)^2}}$ is hence distributed as

$$\chi^2\left(1, \frac{n^2\mu^2}{\left(\frac{\sigma^2}{2\theta}\frac{2e^{-\theta(n-1)\delta} - 2e^{\theta\delta} + n\left(e^{2\theta\delta} - 1\right)}{\left(e^{\theta\delta} - 1\right)^2}\right)^2}\right). \tag{A.10}$$

Acknowledgments

This work was partially supported by the European Union's Horizon 2020 research and innovation program under grant agreement 644732.

References

Balasubramanian, V. (1996). *A geometric formulation of Occam's razor for inference of parametric distributions*. arXiv:9601001.

Barron, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *Annals of Probability*, *13*(4), 1292–1303.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Hoboken, NJ: Wiley.

Fuller, W. A., & Hasza, D. P. (1980). Predictors for the first-order autoregressive process. *Journal of Econometrics*, *13*(2), 139–157.

Fuller, W. A., & Hasza, D. P. (1981). Properties of predictors for autoregressive time series. *Journal of the American Statistical Association*, *76*(373), 155–161.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jeffreys, H. (1998). *The theory of probability*. New York: Oxford University Press.

Jones, L. K. (1990). Constructive approximations for neural networks by sigmoidal functions. *Proceedings of the IEEE*, *78*(10), 1586–1589.

Kuznetsov, V., & Mohri, M. (2015). Learning theory and algorithms for forecasting non-stationary time series. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems, 28* (pp. 541–549). Red Hook, NY: Curran.

McDonald, D. J., Shalizi, C. R., & Schervish, M. (2017). Nonparametric risk bounds for time-series forecasting. *Journal of Machine Learning Research*, *18*(32), 1–40.

Merhav, N., & Feder, M. (1995). A strong version of the redundancy-capacity theorem of universal coding. *IEEE Transactions on Information Theory*, *41*(3), 714–722.

Merhav, N., & Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, *44*(6), 2124–2147.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2018). *Do CIFAR-10 classifiers generalize to CIFAR-10?* arXiv:1806.00451.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, *30*(4), 629–636.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Rissanen, J. (2007). *Information and complexity in statistical modeling*. New York: Springer Science & Business Media.

Vajda, I. (1989). *Theory of statistical inference and information*. Amsterdam: Kluwer Academic.

Vapnik, V. (1998). *Statistical learning theory.* Hoboken, NJ: Wiley.

Vapnik, V. (2013). *The nature of statistical learning theory*. New York: Springer Science & Business Media.

Received November 29, 2018; accepted September 6, 2019.